Thompson sampling for one-dimensional exponential family bandits

Nathaniel Korda, Emilie Kaufmann, Rémi Munos

Thompson Sampling for 1-Dimensional Exponential Family Bandits

Nathaniel Korda, Emilie Kaufmann, and Rémi Munos

Telecom Paristech UMR CNRS 5141 & INRIA Lille - Nord Europe

July 15, 2013

Abstract

Thompson Sampling has been demonstrated in many complex bandit models, however the theoretical guarantees available for the parametric multi-armed bandit are still limited to the Bernoulli case. Here we extend them by proving asymptotic optimality of the algorithm using the Jeffreys prior for 1-dimensional exponential family bandits. Our proof builds on previous work, but also makes extensive use of closed forms for Kullback-Leibler divergence and Fisher information (and thus Jeffreys prior) available in an exponential family. This allow us to give a finite time exponential concentration inequality for posterior distributions on exponential families that may be of interest in its own right. Moreover our analysis covers some distributions for which no optimistic algorithm has yet been proposed, including heavy-tailed exponential families.

1 Introduction

K-armed bandit problems provide an elementary model for exploration-exploitation tradeoffs found at the heart of many online learning problems. In such problems, an agent is presented with K distributions (also called arms, or actions) $\{p_a\}_{a=1}^K$, from which she draws samples interpreted as rewards she wants to maximize. This objective induces a trade-off between choosing to sample a distribution that has already yielded high rewards, and choosing to sample a relatively unexplored distribution at the risk of loosing rewards in the short term. Here we make the assumption that the distributions, p_a, belong to a parametric family of distributions $\mathcal{P} = \{p(\cdot \mid \theta), \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}$. The bandit model is described by a parameter $\theta_0 = (\theta_1, \dots, \theta_K)$ such that $p_a = p(\cdot \mid \theta_a)$. We introduce the mean function $\mu(\theta) = \mathbb{E}_{X \sim p(\cdot|\theta)}[X]$, and the optimal arm $\theta^* = \theta_{a^*}$ where $a^* = \text{argmax}_a\ \mu(\theta_a)$.

An algorithm, ϕ, for a K-armed bandit problem is a (possibly randomised) method for choosing which distribution to sample from next, given a history of previous arm choices and obtained rewards, $\mathcal{H}_{t-1} := ((a_s, x_s))_{s=1}^{t-1}$: each reward x_s is drawn from the distribution p_{a_s}. We denote by ϕ_t the distribution over $\{1, \dots, K\}$ induced by the history $\mathcal{H}_{t-1}$: at time t the agent using ϕ picks arm a with probability $\phi_t(a)$. The agent's goal is to design an algorithm with low regret:

$$\mathcal{R}(\phi, t) = \mathcal{R}(\phi, t)(\theta) := t\mu(\theta^*) - \mathbb{E}_\phi \left[\sum_{s=1}^{t} x_s \right].$$

This quantity measures the expected performance of algorithm ϕ compared to the expected performance of an optimal algorithm given knowledge of the reward distributions, i.e. sampling always from the distribution with the highest expectation.

Since the early 2000s the "optimisim in the face of uncertainty" heuristic has been a popular approach to this problem, providing both simplicity of implementation and finite-time upper bound on the regret (e.g. [4, 7]). However in the last two years there has been renewed interest in the Thompson Sampling heuristic (TS). While this heuristic was first put forward to solve bandit problems eighty years ago in [14], it was not until recently that

theoretical analyses of its performance were achieved [1, 2, 10, 12]. In this paper we take a major step towards generalising these analyses to the same level of generality already achieved for "optimistic" algorithms.

Thompson Sampling Unlike optimistic algorithm which are often based on confidence intervals, the Thompson Sampling algorithm ϕ^{TS,π_0} uses Bayesian tools and puts a prior distribution $\pi_{a,0} = \pi_0$ on each θ_a. A posterior distribution, $\pi_{a,t}$, is then maintained according to the rewards observed in $\mathcal{H}_{t-1}$. At each time a sample $\theta_{a,t}$ is drawn from each posterior $\pi_{a,t}$ and then the algorithm chooses to sample $a_t = \arg\max_{a\in\{1,\dots,K\}}\{\mu(\theta_{a,t})\}$. Therefore $\phi_t^{TS,\pi_0}(a)$ is the posterior probability that $a = a^*$ given the history $\mathcal{H}_{t-1}$.

Our Contributions TS has proved to have impressive empirical performances, very close to those of state of the art algorithms such as DMED and KL-UCB [10, 9, 7]. Furthermore recent works [10, 2] have shown that in the special case where each p_a is a Bernoulli distribution $\mathcal{B}(\theta_a)$, TS using a uniform prior over the arms is asymptotically optimal in the sense that it achieves the asymptotic lower bound on the regret provided by Lai and Robbins in [11] (that holds for univariate parametric bandits). In this paper, we show this optimality property also holds for 1-dimensional exponential families if the algorithm uses the Jeffrey's prior:

Theorem 1. *Suppose that the rewards distributions belong to a* 1*-dimensional canonical exponential family and that π_J is the Jeffrey's prior. Then,*

$$\lim_{T\to\infty} \frac{\mathcal{R}(\phi^{TS,\pi_J}, T)}{\ln T} = \sum_{a=1}^{K} \frac{\mu(\theta_{a^*}) - \mu(\theta_a)}{K(\theta_a, \theta_{a^*})}, \tag{1}$$

where $K(\theta, \theta') := KL(p_\theta, p'_\theta)$ is the Kullback-Leibler divergence between p_θ and p'_θ.

This theorem follows directly from Theorem 2. In the proof of this result we provide in Theorem 4 a finite-time, exponential concentration bound for posterior distributions of exponential family random variables, something that to the best of our knowledge is new to the literature and of interest in its own right. Our proof also exploits the explicit connection between the Jeffreys prior, Fisher information and the Kullback-Leibler divergence in exponential families.

Related Work Another line of recent work has focused on distribution-independent bounds for Thompson Sampling. [2] establishes that $\mathcal{R}(\phi^{TS,\pi_U}, T) = O(\sqrt{KT\ln(T)})$ for Thompson Sampling for bounded rewards (with the classic uniform prior on the underlying Bernoulli parameter). [13] go beyond the Bernoulli model, and give an upper bound on the Bayes risk (i.e. the regret averaged over the prior) independent of the prior distribution. For the parametric multi-armed bandit with K arms described above, their result states that the regret of Thompson Sampling using a prior π_0 is not too big when averaged over this same prior:

$$\mathbb{E}_{\theta\sim\pi_0^{\otimes K}}[\mathcal{R}(\phi^{TS,\pi_0}, T)(\theta)] \leq 4 + K + 4\sqrt{KT\log(T)}.$$

Building on the same ideas, [6] have improved this upper bound to $14\sqrt{KT}$. In our paper, we rather see the prior used by Thompson Sampling as a tool, and we want therefore to obtain garantees for any given problem parametrized by θ.

[13] also use Thompson Sampling in more general models, like the linear bandit model. Their result is a bound on the Bayes risk that does not depend on the prior, whereas Agrawal and Goyal give in [3] a first regret bound for this model. Linear bandits consider a possibly infinite number of arms whose mean rewards are linearly related by a single, unknown coefficient vector. Once again, the analysis in [3] encounters the problem of describing the concentration of posterior distributions. However by using a conjugate normal prior, they can employ explicit the concentration bounds available for Normal distributions to complete their argument.

Paper Structure In Section 2 we describe important features of the one-dimensional canonical exponential families we consider, including closed-form expression for KL-divergences and the Jeffrey's prior. Section 3 gives statements of the main results, and provides the proof of the regret bound. Section 4 proves the posterior concentration result used in the proof of the regret bound.

2 Exponential Families and Jeffreys Priors

A distribution is said to belong to a one-dimensional canonical exponential family if it has a density with respect to some reference measure ν of the form:

$$p(x \mid \theta) = A(x) \exp(T(x)\theta - F(\theta)), \tag{2}$$

where $\theta \in \Theta \subset \mathbb{R}$. T and A are some fixed functions that characterize the exponential family and $F(\theta) = \log\left(\int A(x) \exp\left[T(x)\theta\right] d\lambda(x)\right)$. Θ is called the *parameter space*, $T(x)$ the *sufficient statistic*, and $F(\theta)$ the *normalisation function*. We make the classic assumption that F is twice differentiable with a continuous second derivative. It is well known [16] that:

$$\mathbb{E}_{X|\theta}(T(X)) = F'(\theta) \quad \text{and} \quad \text{Var}_{X|\theta}[T(X)] = F''(\theta)$$

showing in particular that F is strictly convex. The mean function μ is differentiable and stricly increasing, since we can show that

$$\mu'(\theta) = \text{Cov}_{X|\theta}(X, T(X)) > 0.$$

In particular, this shows that μ is one-to-one in θ.

KL-divergence in Exponential Families In an exponential family, a direct computation show that the Kullback-Leibler divergence can be expressed as a Bregman divergence of the normalisation function, F:

$$\text{K}(\theta, \theta') = D_F^B(\theta', \theta) := F(\theta') - \left[F(\theta) + F'(\theta)(\theta' - \theta)\right]. \tag{3}$$

Jeffreys prior in Exponential Families In the Bayesian literature, a special "non-informative" prior, one which is invariant under re-parametrisation of the parameter space, is sometimes considered. It is called the Jeffrey's prior, and it can be shown to be proportional to the square-root of the Fisher information $I(\theta)$. In the special case of the canonical exponential family, the Fisher information takes the form $I(\theta) = F''(\theta)$, hence the Jeffrey's prior for the model (2) is

$$\pi_J(\theta) \propto \sqrt{|F''(\theta)|}.$$

Under the Jeffrey's prior, the posterior on θ after n observations is given by

$$p(\theta|y_1, \dots y_n) \propto \sqrt{F''(\theta)} \exp\left(\theta \sum_{i=1}^{n} T(y_i) - nF(\theta_i)\right) \tag{4}$$

When $\int_\Theta \sqrt{F''(\theta)} d\theta < +\infty$, the prior is called *proper*. However, stasticians often use priors which are not proper: the prior is called *improper* if $\int_\Theta \sqrt{F''(\theta)} d\theta = +\infty$ and any observation makes the corresponding posterior (4) integrable.

Some Intuition for choosing the Jeffreys Prior In the proof of our concentration result for posterior distributions (Theorem 4) it will be crucial to lower bound the prior probability of an ϵ-sized KL-divergence ball around each of the parameters θ_a. Since the Fisher information $F''(\theta) = \lim_{\theta' \to \theta} K(\theta, \theta')/|\theta - \theta'|^2$, choosing a prior proportional to $F''(\theta)$ ensures that the prior measure of such balls are $\Omega(\sqrt{\epsilon})$.

Examples and Pseudocode Algorithm 1 presents pseudocode for Thompson Sampling with the Jeffreys prior for distributions parametrized by their natural parameter θ. But as the Jeffreys prior is invariant under reparametrization, if a distribution is parametrised by some parameter $\lambda \not\equiv \theta$, the algorithm can use the Jeffrey's prior $\propto \sqrt{I(\lambda)}$ on λ, drawing samples from the posterior on λ. Note that the posterior sampling step (in bold) is always tractable using, for example, a Hastings-Metropolis algorithm.

Some examples of common exponential family models are given in Figure 2, together with the posterior distributions on the parameter λ that is used by TS with Jeffreys prior. In addition to examples already studied in [7] for

Name	Distribution	θ	Prior on λ	Posterior on λ
$\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x}\delta_{0,1}$	$\log\left(\frac{\lambda}{1-\lambda}\right)$	Beta $\left(\frac{1}{2},\frac{1}{2}\right)$	Beta $\left(\frac{1}{2}+s,\frac{1}{2}+n-s\right)$
$\mathcal{N}(\lambda,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$	$\frac{\lambda}{\sigma^2}$	$\propto 1$	$\mathcal{N}\left(\frac{s}{n},\frac{\sigma^2}{n}\right)$
$\Gamma(k,\lambda)$	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}1_{[0,+\infty[}(x)$	$-\lambda$	$\propto \frac{1}{\lambda}$	$\Gamma(kn,s)$
$\mathcal{P}(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}\delta_{\mathbb{N}}$	$\log(\lambda)$	$\propto \frac{1}{\sqrt{\lambda}}$	$\Gamma\left(\frac{1}{2}+s,n\right)$
Pareto(x_m,λ)	$\frac{\lambda x_m^\lambda}{x^{\lambda+1}}1_{[x_m,+\infty[}(x)$	$-\lambda-1$	$\propto \frac{1}{\lambda}$	$\Gamma\left(n+1,s-n\log x_m\right)$
Weibull(k,λ)	$k\lambda(x\lambda)^{k-1}e^{-(\lambda x)^k}1_{[0,+\infty[}$	$-\lambda^k$	$\propto \frac{1}{\lambda^k}$	$\alpha\lambda^{(n-1)k}\exp(-\lambda^k s)$

Figure 1: The posterior distribution after observations $y_1,\dots,y_n$ depends on n and $s=\sum_{i=1}^n T(y_i)$

Algorithm 1 Thompson Sampling for Exponential Families with Jeffrey's prior

Require: F normalization function, T sufficient statistic, μ mean function
  for $t=1\dots K$ **do**
    Sample arm t and get rewards x_t
    $N_t=1$, $S_t=T(x_t)$.
  end for
  for $t=K+1\dots n$ **do**
    for $a=1\dots K$ **do**
      Sample $\theta_{a,t}$ **from** $\pi_{a,t}\propto\sqrt{F''(\theta)}\exp\left(\theta S_a-N_aF(\theta)\right)$
    end for
    Sample arm $A_t=\text{argmax}_a\mu(\theta_{a,t})$ and get reward x_t
    $S_{A_t}=S_{A_t}+T(x_t)$ $N_{A_t}=N_{A_t}+1$
  end for

which $T(x)=x$, we also give two examples of more general canonical exponential families, namely the Pareto distribution with known min value and unknown tail index λ, Pareto(x_m,λ), for which $T(x)=\log(x)$, and the Weibul distribution with known shape and unknown rate parameter, Weibull(k,λ), for which $T(x)=x^k$. These last two distributions are not covered even by the work in [8], and belong to the family of heavy-tailed distributions.

For the Bernoulli model, one note futher that the use of the Jeffreys prior is not covered by the previous analyses. These analyses make an extensive use of the uniform prior, through the fact that the coefficient of the Beta posteriors they consider have to be integers.

3 Results and Proof of Regret Bound

An *exponential family K-armed bandit* is a K-armed bandit for which the reward distributions p_a are known to be elements of an exponential family of distributions $\mathcal{P}(\Theta)$. We denote by p_{θ_a} the distribution of arm a and its mean by $\mu_a=\mu(\theta_a)$.

Theorem 2 (Regret Bound). *Assume that $\mu_1>\mu_a$ for all $a\neq 1$, and that $\pi_{a,0}$ is taken to be the Jeffrey's prior over Θ. Then for every $\epsilon>0$ there exists a constant $\mathcal{C}(\epsilon,\mathcal{P})$ depending on ϵ and on the problem $\mathcal{P}$ such that the regret of Thompson Sampling using the Jeffrey's prior satisfies*

$$\mathcal{R}(\phi^{TS,\pi_J},T)\leq\frac{1+\epsilon}{1-\epsilon}\left(\sum_{a=2}^{K}\frac{(\mu_1-\mu_a)}{K(\theta_a,\theta_1)}\right)\ln(T)+\mathcal{C}(\epsilon,\mathcal{P}).$$

Proof: We give here the main argument of the proof of the regret bound, which proceed by bounding the expected number of draws of any suboptimal arm. Along the way we shall state concentration results whose proofs are

postponed to later sections.

Step 0: Notation We denote by $y_{a,s}$ the s-th observation of arm a and by $N_{a,t}$ the number of times arm a is chosen up to time t. $(y_{a,s})_{s\geq 1}$ is i.i.d. with distribution p_{θ_a}. Let $Y_a^u := (y_{a,s})_{1\leq s\leq u}$ be the vector of first u observations from arm a. $Y_{a,t} := Y_a^{N_{a,t}}$ is therefore the vector of observations from arm a available at the beginning of round t. Recall that $\pi_{a,t}$, respectively $\pi_{a,0}$, is the posterior, respectively the prior, on θ_a at round t of the algorithm.

We let $L(\theta) := \frac{1}{2}\min(\sup_y p(y|\theta), 1)$. For any $\delta_a > 0$, we introduce the event $\tilde{E}_{a,t} = \tilde{E}_{a,t}(\delta_a)$:

$$\tilde{E}_{a,t} = \left(\exists 1 \leq s' \leq N_{a,t} : p(y_{a,s'}|\theta_a) \geq L(\theta_a), \left|\frac{\sum_{s=1,s\neq s'}^{N_{a,t}} T(y_{a,s})}{N_{a,t}-1} - F'(\theta_a)\right| \leq \delta_a\right). \tag{5}$$

For all $a \neq 1$ and Δ_a such that $\mu_a < \mu_a + \Delta_a < \mu_1$, we introduce

$$E_{a,t}^{\theta} = E_{a,t}^{\theta}(\Delta_a) := \big(\mu\left(\theta_{a,t}\right) \leq \mu_a + \Delta_a\big).$$

On $\tilde{E}_{a,t}$, the empirical sufficient statistic of arm a at round t is well concentrated around its mean and a 'likely' realization of arm a has been observed. On $E_{a,t}^{\theta}$, the mean of the distribution with parameter $\theta_{a,t}$ does not exceed by much the true mean, μ_a. δ_a and Δ_a will be carefully chosen at the end of the proof.

Step 1: Concentration Results We state here the two concentration results that are necessary to evaluate the probability of the above events.

Lemma 3. *Let (y_s) be an i.i.d sequence of distribution $p(\cdot \mid \theta)$ and $\delta > 0$. Then*

$$\mathbb{P}\left(\left|\frac{1}{u}\sum_{s=1}^{u}[T(y_s) - F'(\theta)]\right| \geq \delta\right) \leq 2e^{-u\tilde{K}(\theta,\delta)},$$

where $\tilde{K}(\theta,\delta) = \min(K(\theta + g(\delta), \theta), K(\theta - h(\delta), \theta))$, with $g(\delta) > 0$ defined by $F'(\theta + g(\delta)) = F'(\theta) + \delta$ and $h(\delta) > 0$ defined by $F'(\theta - h(\delta)) = F'(\theta) - \delta$.

The two following inequalities that will be useful in the sequel can easily be deduced from Lemma 3. Their proof is gathered in Appendix A with that of Lemma 3. For any arm a,

$$\sum_{t=1}^{T}\mathbb{P}(a_t = a, \tilde{E}_{a,t}(\delta_a)^c) \leq \sum_{t=1}^{\infty}\mathbb{P}\left(p(y_{a,1}|\theta_a) \leq L(\theta_a)\right)^t + \sum_{t=1}^{\infty} 2te^{-(t-1)\tilde{K}(\theta_a,\delta_a)} \tag{6}$$

$$\sum_{t=1}^{T}\mathbb{P}(\tilde{E}_{a,t}(\delta_a)^c | N_{a,t} > t^b) \leq \sum_{t=1}^{\infty}\mathbb{P}\left(p(y_{a,1}|\theta_a) \leq L(\theta_a)\right)^{t^b} + \sum_{t=1}^{\infty} 2t^2e^{-(t^b-1)\tilde{K}(\theta_a,\delta_a)} \tag{7}$$

The second result tells us that concentration of the empirical sufficient statistic around its mean implies concentration of the posterior distribution around the true parameter:

Theorem 4 (Posterior Concentration). *Let $\pi_{a,0}$ be the Jeffreys' prior. There exists constants $C_{1,a} = C_1(F,\theta_a) > 0$, $C_{2,a} = C_2(F,\theta_a,\Delta_a) > 0$, and $N(\theta_a, F)$ s.t., $\forall N_{a,t} \geq N(\theta_a, F)$,*

$$\mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}\big(\mu(\theta_{a,t}) > \mu(\theta_a) + \Delta_a | Y_{a,t}\big) \leq C_{1,a}e^{-(N_{a,t}-1)(1-\delta_a C_{2,a})K(\theta_a,\mu^{-1}(\mu_a+\Delta_a))+\ln(N_{a,t})}$$

whenever $\delta_a < 1$ and Δ_a are such that $1 - \delta_a C_{2,a}(\Delta_a) > 0$.

Step 2: Lower Bound the Number of Optimal Arm Plays with High Probability The main difficulty adressed in previous regret analyses for Thompson Sampling is the control of the number of draws of the optimal arm. We provide this control in the form of Proposition 5 which is adapted from Proposition 1 in [10] whose proof, an outline of which is given in Appendix D, explores in depth the randomised nature of Thompson Sampling. In particular, we show that the proof in [10] can be significantly simplified, but at the expense of no longer being able to describe the constant C_b explicitly:

Proposition 5. *For any $b \in (0,1)$ there exists a constant $C_b(\pi, \mu_1, \mu_2, K) < \infty$ such that*

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_{1,t} \leq t^b\right) \leq C_b.$$

Step 3: Decomposition The idea in this step is to decompose the probability of playing a suboptimal arm into principle and negligible components and control these components with the results from Steps 1 and 2:

$$\sum_{t=1}^{T} \mathbb{P}\left(a_t = a\right) = \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, E^{\theta}_{a,t}\right)}_{(A)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, (E^{\theta}_{a,t})^c\right)}_{(B)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}^c_{a,t}\right)}_{(C)}. \tag{8}$$

The terms (B) and (C) are about concentration of the posterior on the suboptimal arm. An upper bound on term (C) is given in (6), whereas a bound on term (B) follows from Lemma 6 below. Although the proof of this lemma is standard, and bears a strong similarity to Lemma 3 of [3], we provide it in Appendix C for the sake of completeness.

Lemma 6. *For all actions a and for all $\epsilon > 0$, $\exists\, N_\epsilon = N_\epsilon(\delta_a, \Delta_a, \theta_a) > 0$ such that*

$$(B) \leq [(1-\epsilon)(1-\delta_a C_{2,a}) K(\theta_a, \mu^{-1}(\mu_a + \Delta_a))]^{-1} \ln(T) + \max\{N_\epsilon, N(\theta_a, F)\} + 1.$$

where $N_\epsilon = N_\epsilon(\delta_a, \Delta_a, \theta_a)$ is the smallest integer such that for all $n \geq N_\epsilon$

$$(n-1)^{-1} \ln(C_{1,a} n) < \epsilon (1 - \delta_a C_{2,a}) K(\theta_a, \mu^{-1}(\mu_a + \Delta_a)),$$

and $N(\theta_a, F)$ is the constant from Theorem 4.

When we have seen enough observations on the optimal arm, term (A) also becomes a result about the concentration of the posterior, but this time for the optimal arm:

$$\begin{aligned}(A) &\leq \sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, E^{\theta}_{a,t} \mid N_{1,t} > t^b\right) + C_b \leq \sum_{t=1}^{T} \mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta'_a \mid N_{1,t} > t^b\right) + C_b \\ &\leq \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta'_a, \tilde{E}_{1,t}(\delta_1) \mid N_{1,t} > t^b\right)}_{B'} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(\tilde{E}^c_{1,t}(\delta_1) \mid N_{1,t} > t^b\right)}_{C'} + C_b \end{aligned} \tag{9}$$

where $\Delta'_a = \mu_1 - \mu_a - \Delta_a$ and $\delta_1 > 0$ remains to be chosen. The first inequality comes from Proposition 5, and the second inequality comes from the following fact: if arm 1 is not chosen and arm a is such that $\mu(\theta_{a,t}) \leq \mu_a + \Delta_a$, then $\mu(\theta_{1,t}) \leq \mu_a + \Delta_a$. A bound on term (C') is given in (7) for $a = 1$ and δ_1. In Theorem 4, we bound the conditional probability that $\mu(\theta_{a,t})$ exceed the true mean. Following the same lines, we can also show that, on $\tilde{E}_{1,t}(\delta_1)$,

$$\mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta'_a | Y_{1,t}\right) \leq C_{1,1} e^{-(N_{1,t}-1)(1-\delta_1 C_{2,1}) \mathrm{K}(\theta_1, \mu^{-1}(\mu_1 - \Delta'_a)) + \ln(N_{1,t})}.$$

For any $\Delta'_a > 0$, one can choose δ_1 such that $1 - \delta_1 C_{1,1} > 0$. Then, with $N = N(\mathcal{P})$ such that the function

$$u \mapsto e^{-(u-1)(1-\delta_1 C_{2,1}) \mathrm{K}(\theta_1, \mu^{-1}(\mu_1 - \Delta'_a)) + \ln u}$$

is decreasing for $u \geq N$, (B') is bounded by

$$N^{1/b} + \sum_{t=N^{1/b}+1}^{\infty} C_{1,1} e^{-(t^b-1)(1-\delta_1 C_{2,1}) \mathrm{K}(\theta_1, \mu^{-1}(\mu_1 - \Delta'_a)) + \ln(t^b)} < \infty.$$

Step 4: Choosing the Values δ_a and ϵ_a So far, we have shown that for any $\epsilon > 0$ and for any choice of $\delta_a > 0$ and $0 < \Delta_a < \mu_1 - \mu_a$ such that $1 - \delta_a C_{2,a} > 0$, there exists a constant $\mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P})$ such that

$$\mathbb{E}[N_{a,T}] \leq \frac{\ln(T)}{(1-\delta_a C_{2,a}) K(\theta_a, \mu^{-1}(\mu_a + \Delta_a))(1-\epsilon)} + \mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P})$$

The constant is of course increasing (dramatically) when δ_a goes to zero, Δ_a to $\mu_1 - \mu_a$, or ϵ to zero. But one can choose Δ_a close enough to $\mu_1 - \mu_a$ and δ_a small enough, such that

$$(1 - C_{2,a}(\Delta_a)\delta_a)\mathrm{K}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)) \geq \frac{\mathrm{K}(\theta_a, \theta_1)}{(1+\epsilon)},$$

and this choice leads to

$$\mathbb{E}[N_{a,T}] \leq \frac{1+\epsilon}{1-\epsilon} \frac{\ln(T)}{\mathrm{K}(\theta_a, \theta_1)} + \mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P}).$$

Using that $\mathcal{R}(\phi, T) = \sum_{a=2}^{K} (\mu_1 - \mu_a)\mathbb{E}[N_{a,T}]$ concludes the proof. □

4 Posterior Concentration: Proof of Theorem 4

For ease of notation, we drop the subscript a and let (y_s) be an i.i.d. sequence of distribution p_θ, with mean $\mu = \mu(\theta)$. Furthermore, by conditioning on the value of N_s, it is enough to bound $\mathbf{1}_{\tilde{E}_u} \mathbb{P}\left(\mu(\theta_u) \geq \mu + \Delta | Y^u\right)$ where $Y^u = (y_s)_{1 \leq s \leq u}$ and

$$\tilde{E}_u = \left(\exists 1 \leq s' \leq u : p(y_{s'}|\theta) \geq L(\theta), \left| \frac{\sum_{s=1, s \neq s'}^{u} T(y_s)}{u-1} - F'(\theta) \right| \leq \delta \right).$$

Step 1: Extracting a Kullback-Leibler Rate The argument rests on the following Lemma, whose proof can be found in Appendix B

Lemma 7. *Let $\tilde{E}_u$ be the event defined by* (5)*, and introduce $\Theta_{\theta,\Delta} := \{\theta' \in \Theta : \mu(\theta') \geq \mu(\theta) + \Delta\}$. The following inequality holds:*

$$\mathbf{1}_{\tilde{E}_u} \mathbb{P}\left(\mu(\theta_u) \geq \mu + \Delta | Y^u\right) \leq \frac{\int_{\theta' \in \Theta_{\theta,\Delta}} e^{-(u-1)\left(K[\theta,\theta'] - \delta|\theta - \theta'|\right)} \pi(\theta'|y_{s'}) d\theta'}{\int_{\theta' \in \Theta} e^{-(u-1)(K[\theta,\theta'] + \delta|\theta - \theta'|)} \pi(\theta'|y_{s'}) d\theta'}, \tag{10}$$

with $s' = \inf\{s \in \mathbb{N} : p(y_s|\theta) \geq L\}$.

Step 2: Upper bounding the numerator of (10) We first note that on $\Theta_{\theta,\Delta}$ the leading term in the exponential is $K(\theta, \theta')$. Indeed, from (3) we know that

$$K(\theta, \theta')/|\theta - \theta'| = |F'(\theta) - (F(\theta) - F(\theta'))/(\theta - \theta')|$$

which, by strict convexity of F, is strictly increasing in $|\theta - \theta'|$ for any fixed θ. Now since μ is one-to-one and continuous, $\Theta^c_{\theta,\Delta}$ is an interval whose interior contains θ, and hence, on $\Theta_{\theta,\Delta}$,

$$\frac{K(\theta, \theta')}{|\theta - \theta'|} \geq \frac{F(\mu^{-1}(\mu + \Delta)) - F(\theta)}{\mu^{-1}(\mu + \Delta) - \theta} - F'(\theta) := (C_2(F, \theta, \Delta))^{-1} > 0.$$

So for δ such that $1 - \delta C_2 > 0$ we can bound the numerator of (10) by:

$$\begin{aligned}
&\int_{\theta' \in \Theta_{\theta,\Delta}} e^{-(u-1)(K(\theta,\theta') - \delta|\theta - \theta'|)} \pi(\theta'|y_{s'}) d\theta' \leq \int_{\theta' \in \Theta_{\theta,\Delta}} e^{-(u-1)K(\theta,\theta')(1-\delta C_2)} \pi(\theta'|y_{s'}) d\theta' \\
&\quad \leq e^{-(u-1)(1-\delta C_2)\mathrm{K}(\theta, \mu^{-1}(\mu+\Delta))} \int_{\Theta_{\theta,\Delta}} \pi(\theta'|y_{s'}) d\theta' \leq e^{-(u-1)(1-\delta C_2)\mathrm{K}(\theta, \mu^{-1}(\mu+\Delta))}
\end{aligned} \tag{11}$$

where we have used that $\pi(\cdot|y_{s'})$ is a probability distribution, and that, since μ is increasing, $\mathrm{K}(\theta, \mu^{-1}(\mu + \Delta)) = \inf_{\theta' \in \Theta_{\theta,\Delta}} K(\theta, \theta')$.

Step 3: Lower bounding the denominator of (10) To lower bound the denominator, we reduce the integral on the whole space Θ to a KL-ball, and use the structure of the prior to lower bound the measure of that KL-ball under the posterior obtained with the well-chosen observation $y_{s'}$. We introduce the following notation for KL balls: for any $x \in \Theta,\ \epsilon > 0$, we define

$$B_\epsilon(x) := \{\theta' \in \Theta : K(x, \theta') \leq \epsilon\}.$$

We have $\frac{K(\theta,\theta')}{(\theta-\theta')^2} \to F''(\theta) \neq 0$ (since F is strictly convex). Therefore, there exists $N_1(\theta, F)$ such that for $u \geq N_1(\theta, F)$, on $B_{\frac{1}{u^2}}(\theta)$,

$$|\theta - \theta'| \leq \sqrt{2K(\theta, \theta')/F''(\theta)}.$$

Using this inequality we can then bound the denominator of (10) whenever $u \geq N_1(\theta, F)$ and $\delta < 1$:

$$\int_{\theta' \in \Theta} e^{-(u-1)(K(\theta,\theta')+\delta|\theta-\theta'|)}\pi(\theta'|y_{s'})d\theta' \geq \int_{\theta' \in B_{1/u^2}(\theta)} e^{-(u-1)(K(\theta,\theta')+\delta|\theta-\theta'|)}\pi(\theta'|y_{s'})d\theta'$$

$$\geq \int_{\theta' \in B_{1/u^2}(\theta)} e^{-(u-1)\left(K(\theta,\theta')+\delta\sqrt{\frac{2K(\theta,\theta')}{F''(\theta)}}\right)}\pi(\theta'|y_{s'})d\theta' \geq \pi\left(B_{1/u^2}(\theta)|y_{s'}\right) e^{-\left(1+\sqrt{\frac{2}{F''(\theta)}}\right)}. \tag{12}$$

Finally we turn our attention to the quantity

$$\pi\left(B_{1/u^2}(\theta)|y_{s'}\right) = \frac{\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\pi_0(\theta')d\theta'}{\int_\Theta p(y'_s|\theta')\pi_0(\theta')d\theta'} = \frac{\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\sqrt{F''(\theta')}d\theta'}{\int_\Theta p(y'_s|\theta')\sqrt{F''(\theta')}d\theta'}. \tag{13}$$

Now since the KL divergence is convex in the second argument, we can write $B_{1/u}(\theta) = (a, b)$. So, from the convexity of F we deduce that

$$\frac{1}{u^2} = K(\theta, b) = F(b) - [F(\theta) + (b-\theta)F'(\theta)] = (b-\theta)\left[\frac{F(b)-F(\theta)}{(b-\theta)} - F'(\theta)\right]$$
$$\leq (b-\theta)\left[F'(b) - F'(\theta)\right] \leq (b-a)\left[F'(b) - F'(\theta)\right] \leq (b-a)\left[F'(b) - F'(a)\right].$$

As $p(y \mid \theta) \to 0$ as $y \to \pm\infty$, the set $\mathcal{C}(\theta) = \{y : p(y \mid \theta) \geq L(\theta)\}$ is compact. The map $y \mapsto \int_\Theta p(y|\theta')\sqrt{F''(\theta')}d\theta' < \infty$ is continuous on the compact $\mathcal{C}(\theta)$. Thus, it follows that

$$L'(\theta) = L'(\theta, F) := \sup_{y : p(y|\theta) > L(\theta)} \left\{\int_\Theta p(y|\theta')\sqrt{F''(\theta')}d\theta'\right\} < \infty$$

is an upper bound on the denominator of (13).

Now by the continuity of F'', and the continuity of $(y, \theta) \mapsto p(y|\theta)$ in both coordinates, there exists an $N_2(\theta, F)$ such that for all $u \geq N_2(\theta, F)$

$$F''(\theta) \geq \frac{1}{2}\frac{F'(b) - F'(a)}{b-a} \text{ and } \left(p(y|\theta')\sqrt{F''(\theta')} \geq \frac{L(\theta)}{2}\sqrt{F''(\theta)},\ \forall \theta' \in B_{1/u^2}(\theta), y \in \mathcal{C}(\theta)\right).$$

Finally, for $u \geq N_2(\theta, F)$, we have a lower bound on the numerator of (13):

$$\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\sqrt{F''(\theta')}d\theta' \geq \frac{L(\theta)}{2}\sqrt{F''(\theta)}\int_a^b d\theta' = \frac{L(\theta)}{2}\sqrt{(F'(b) - F'(a))(b-a)} \geq \frac{L(\theta)}{2u}$$

Puting everything together, we get that there exist constants $C_2 = C_2(F, \theta, \Delta)$ and $N(\theta, F) = \max\{N_1, N_2\}$ such that for every $\delta < 1$ satisfying $1 - \delta C_2 > 0$, and for every $u \geq N$, one has

$$\mathbf{1}_{\tilde{E}_u}\mathbb{P}(\mu(\theta_u) \geq \mu(\theta) + \Delta|Y_u) \leq \frac{2e^{1+\sqrt{\frac{2}{F''(\theta)}}}L'(\theta)u}{L(\theta)}e^{-(u-1)(1-\delta C_2)\mathrm{K}(\theta,\mu^{-1}(\mu+\Delta))}.$$

Remark 8. *Note that when the prior is proper we do not need to introduce the observation $y_{s'}$, which significantly simplifies the argument. Indeed in this case, in (11) we can use π_0 in place of $\pi(\cdot|y_{s'})$ which is already a probability distribution. In particular, the quantity* (13) *is replaced by $\pi_0\left(B_{1/u^2}(\theta)\right)$, and so the constants L and L' are not needed.*

5 Conclusions

We have shown that choosing to use the Jeffrey's prior in Thompson Sampling leads to an asymptotically optimal algorithm for bandit models whose rewards belong to a 1-dimensional canonical exponential family. The cornerstone of our proof is a finite time concentration bound for posterior distributions in exponential families, which, to the best of our knowledge, is new to the literature. With this result we built on previous analyses and avoided Bernoulli-specific arguments. Thompson Sampling with Jeffreys prior is now a provably competitive alternative to KL-UCB for exponential family bandits. Moreover our proof holds for slightly more general problems than those for which KL-UCB is provably optimal, including some heavy-tailed exponential family bandits.

Our arguments are potentially generalisable. Notably generalising to n-dimensional exponential family bandits requires only generalising Lemma 3 and Step 3 in the proof of Theorem 4. Our result is asymptotic, but the only stage where the constants are not explicitly derivable from knowledge of F, T, and θ_0 is in Lemma 9. Future work will investigate these open problems. Another possible future direction lies the optimal choice of prior distribution. Our theoretical guarantees only hold for Jeffreys prior, but a careful examination of our proof shows that the important property is to have, for every θ_a,

$$-\ln\left(\int_{(\theta':\mathrm{K}(\theta_a,\theta')\leq n^{-2})}\pi_0(\theta')d\theta'\right)=o\left(n\right),$$

which could hold for prior distributions other than the Jeffreys prior.

References

[1] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference On Learning Theory (COLT)*, 2012.

[2] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[3] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *30th International Conference on Machine Learning (ICML)*, 2013.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities*. Oxford Univeristy Press, 2013.

[6] S. Bubeck and Che-Yu Liu. A note on the bayesian regret of thompson sampling with an arbitrairy prior. arXiv:1304.5758, 2013.

[7] O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *to appear in Annals of Statistics*, 2013.

[8] A. Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Conference On Learning Theory (COLT)*, 2011.

[9] J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *Conference On Learning Theory (COLT)*, 2010.

[10] E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 199–213. Springer, 2012.

[11] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[12] B.C. May, N. Korda, A. Lee, and D. Leslie. Optimistic bayesian sampling in contextual bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012.

[13] D. Russo and B. Van Roy. Learning to optimize via posterior sampling. arXiv:1301.2609, 2013.

[14] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[15] A.W Van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

[16] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Publishing Company, Incorporated, 2010.

A Concentration of the Sufficient Statistics: Proof of Lemma 3, and Inequalities (6) and (7)

Proof of Lemma 3. The proof of Lemma 3 follows from the classical Cramér-Chenoff technique (see [5]). For any $\lambda > 0$.

$$\begin{aligned} A :=& \mathbb{P}\left(\frac{1}{u}\sum_{i=1}^{u}[T(y_i) - F'(\theta)] \geq \delta\right) = \mathbb{P}\left(e^{\lambda\left(\sum_{i=1}^{u}[T(y_i)-F'(\theta)]\right)} \geq e^{\lambda u}\delta\right) \\ \leq & e^{-\lambda u\delta}\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{u}[T(y_i)-F'(\theta)]\right)}\right] = e^{-u(\delta\lambda-\phi_a(\lambda))} \end{aligned}$$

where we have used the Markov inequality, and where

$$\phi_a(\lambda) := \ln \mathbb{E}_{X|\theta}\left[e^{\lambda(T(X)-F'(\theta))}\right] = F(\theta+\lambda) - F(\theta) - \lambda F'(\theta).$$

Now we optimize in λ by choosing $\lambda > 0$ that maximizes

$$\delta\lambda - \phi_a(\lambda) = \lambda(\delta + F'(\theta)) - F(\theta+\lambda) + F(\theta) := f(\lambda).$$

$f(\lambda)$ is differentiable in λ and its minimum, λ^*, satisfies $f'(\lambda^*) = 0$ i.e.

$$F'(\theta+\lambda^*) = \delta + F'(\theta).$$

(Note that $\lambda^* > 0$ since F' is increasing). Finally, we get

$$A \leq e^{-u((\delta+F'(\theta))\lambda^* - F(\theta+\lambda^*)+F(\theta))} = e^{-u(F'(\theta+\lambda^*)\lambda^* - F(\theta+\lambda^*)+F(\theta))} = e^{-uK(\theta+\lambda^*,\theta)}.$$

The same reasoning leads to the upper bound

$$\mathbb{P}\left(\frac{1}{u}\sum_{s=1}^{u}[T(y_s) - F'(\theta)] \leq -\delta\right) \leq e^{-u\mathrm{K}(\theta-\nu^*,\theta)},$$

where ν^* is such that $F'(\theta-\nu^*) = F'(\theta) - \delta$. □

For the proof of inequalities (6) and (7), we intoduce the notation $Y_{a,s'}^{u} = Y_a^s \backslash \{y_{a,s}\}$ (the first u observations of arms a exept observation $y_{a,s'}$). First note that we have $\tilde{E}_{a,t}^c \subseteq B_{a,N_{a,t}} \bigcup D_{a,N_{a,t}}$, with

$$\begin{aligned} B_{a,s} &= (\forall s' \in [1,s], p(y_{a,s'}|\theta_a) \leq L(\theta_a)) \\ D_{a,s} &= \left(\exists s' \in \{1,\dots s\} : \left|\frac{1}{s-1}\sum_{k=1,k\neq s'}^{s}(T(y_{a,k}) - F'(\theta_a))\right| \geq \delta_a\right) \end{aligned}$$

One then has

$$\begin{aligned} \sum_{t=1}^{T}\mathbb{P}(a_t = a, \tilde{E}_{a,t}^c(\delta)) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{t}\mathbf{1}_{(a_t=a,N_{a,t}=s)}(\mathbf{1}_{B_{a,s}} + \mathbf{1}_{D_{a,s}})\right] \\ &\leq \mathbb{E}\left[\sum_{s=1}^{T}\mathbf{1}_{B_{a,s}}\right] + \mathbb{E}\left[\sum_{s=1}^{T}\mathbf{1}_{D_{a,s}}\right] \\ &\leq \sum_{s=1}^{T}\mathbb{P}\left(p(y_{a,s'}|\theta_a) \leq L(\theta_a)\right)^s + \sum_{s=1}^{T}\sum_{s'=1}^{s}\mathbb{P}(E_{Y_{a,s'}^s}(\delta_a)^c) \\ &\leq \sum_{s=1}^{\infty}\mathbb{P}\left(p(y_{a,s'}|\theta_a) \leq L(\theta_a)\right)^s + \sum_{s=1}^{\infty}se^{-(s-1)\tilde{K}(\theta_a,\delta_a)}, \end{aligned}$$

which gives inequality (6). To proof (7), we write:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{P}(\tilde{E}_{a,t}(\delta_a)^c|N_{a,t}>t^b) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=t^b}^{t}\mathbf{1}_{N_{a,t}=s}(\mathbf{1}_{B_{a,s}}+\mathbf{1}_{D_{a,s}})\right]\\ &\leq \sum_{t=1}^{T}\sum_{s=t^b}^{t}\mathbb{P}(p(y_{a,s'}|\theta_a)\leq L(\theta_a))^s+\sum_{t=1}^{T}\sum_{s=t^b}^{t}\sum_{s'=1}^{s}\mathbb{P}(E_{Y^s_{a,s'}}(\delta_a)^c)\\ &\leq \sum_{t=1}^{T}t\mathbb{P}(p(y_{a,s'}|\theta_a)\leq L(\theta_a))^{t^b}+\sum_{t=1}^{T}t^2\exp(-t^b\tilde{K}(\theta_a,\delta)).\end{aligned}$$

B Extracting the KL-divergence: Proof of Lemma 7

If we assume that the event $\tilde{E}_u$ holds, $s'\leq u$. So, on this event we have

$$\begin{aligned}\mathbb{P}\left(\mu(\theta_u)\geq\mu+\Delta|Y^u\right) &= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}\prod_{s=1,s\neq s'}^{u}p(y_s\mid\theta')p(y_{s'}|\theta')\pi(\theta')d\theta'}{\int_{\theta'\in\Theta}\prod_{s=1,s\neq s'}^{u}p(y_s\mid\theta')p(y_{s'}|\theta')\pi(\theta')d\theta'}\\ &= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}\prod_{s=1,s\neq s'}^{u}\frac{p(y_s|\theta')}{p(y_s|\theta)}p(y_{s'}|\theta')\pi(\theta')d\theta'}{\int_{\theta'\in\Theta}\prod_{s=1,s\neq s'}^{u}\frac{p(y_s|\theta')}{p(y_s|\theta)}p(y_{s'}|\theta')\pi(\theta')d\theta'}\\ &= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}e^{-(u-1)K[Y'^u,\theta,\theta']}\pi(\theta'|y_{s'})d\theta'}{\int_{\theta'\in\Theta}e^{-(u-1)K[Y'^u,\theta,\theta']}\pi(\theta'|y_{s'})d\theta'}\end{aligned}$$

where $\pi(\theta|y_{s'})$ denotes the posterior distribution on θ after observation $y_{s'}$ and

$$K[Y^u_{s'},\theta,\theta']:=\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}\ln\frac{p(y_s\mid\theta)}{p(y_s\mid\theta')}$$

denotes the empirical KL-divergence obtained from the observations $Y^u_{s'}=Y^u\setminus\{y_{s'}\}$. Introducing

$$r(Y^u_{s'},\theta')=K[Y^u_{s'},\theta,\theta']-\mathbb{E}_{X|\theta}\left(\ln\frac{p(X\mid\theta)}{p(X\mid\theta')}\right),$$

we can rewrite

$$\mathbb{P}\left(\mu(\theta_u)\geq\mu+\Delta|Y^u\right)=\frac{\int_{\theta'\in\Theta_{\theta,\Delta}}e^{-(u-1)\left(K[\theta,\theta']+r(Y'^u,\theta')\right)}\pi(\theta'|y_{s'})d\theta'}{\int_{\theta'\in\Theta}e^{-(u-1)(K[\theta,\theta']+r(Y'^u,\theta'))}\pi(\theta'|y_{s'})d\theta'}.$$

Now, a direct computation show that

$$|r(Y'^u,\theta')|\leq|\theta-\theta'|\left|\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}[T(y_s)-F'(\theta)]\right|. \tag{14}$$

Indeed, for that for any $\theta,\theta'\in\Theta$

$$\ln\frac{p(y\mid\theta)}{p(y\mid\theta')}=T(x)(\theta-\theta')-[F(\theta)-F(\theta')],$$

and one also recalls that

$$K(\theta,\theta') = F'(\theta)(\theta-\theta') - [F(\theta) - F(\theta')]. \tag{15}$$

Hence

$$\begin{aligned}|r(Y_{s'}^u,\theta,\theta')| &= \left|\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}\left[\ln\frac{p(y_s\mid\theta)}{p(y_s\mid\theta')} - K(\theta,\theta')\right]\right| \\ &= \left|\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}[(T(x)-F'(\theta))(\theta-\theta')]\right| \leq \left|\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}[T(y_s)-\nabla F(\theta)]\right||\theta'-\theta|.\end{aligned}$$

The inequality (14) leads to the result, using that on $\tilde{E}_u$,

$$\left|\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}[T(y_s)-F'(\theta)]\right| \leq \delta$$

C Proof of Lemma 6

From Theorem 4 we know that, for $N_{a,t} \geq N(\theta_a, F)$,

$$\begin{aligned}\mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}((E_{a,t}^{\theta})^c \mid \mathcal{F}_t) &= \mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}((E_{a,t}^{\theta})^c \mid Y_{a,t}) \\ &\leq C_{1,a}e^{-(N_{a,t}-1)(1-\delta_a C_{2,a})\mathrm{K}(\theta_a,\mu^{-1}(\mu_a+\Delta_a))+\ln N_{a,t}} \\ &\leq e^{-(N_{a,t}-1)\left((1-\delta_a C_{2,a})\mathrm{K}(\theta_a,\mu^{-1}(\mu_a+\Delta_a))-\ln(C_{1,a}N_{a,t})/(N_{a,t}-1)\right)}\end{aligned}$$

Let $N_\epsilon = N_\epsilon(\delta_a,\Delta_a,\theta_a)$ be the smallest integer such that for all $n \geq N_\epsilon$

$$\frac{\ln(C_{1,a}n)}{n-1} < \epsilon(1-\delta_a C_{2,a})\mathrm{K}(\theta_a,\mu^{-1}(\mu_a+\Delta_a)).$$

Defining

$$L_T := \frac{\ln T}{(1-\epsilon)(1-\delta_a C_{2,a})\mathrm{K}(\theta_a,\mu^{-1}(\mu_a+\Delta_a))}$$

we have that for all t and T such that $N_{a,t}-1 \geq \max(L_T, N_\epsilon, N(\theta_a,F))$,

$$\mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}(\mu(\theta_a(t) > \mu(\theta_a)+\Delta_a \mid \mathcal{F}_t) \leq \frac{1}{T}.$$

Let $\tau = \inf\{t\in\mathbb{N} \mid N_{a,t} \geq \max(L_T, N_\epsilon, N(\theta_a,F))+1\}$. τ is a stopping time with respect to $\mathcal{F}_t$. Then,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{P}\left(a_t=a,(E_{a,t}^{\theta})^c,\tilde{E}_{a,t}\right) &\leq \mathbb{E}\left[\sum_{t=1}^{\tau}\mathbf{1}_{(a_t=a)}\right] + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\mathbf{1}_{(a_t=a)}\mathbf{1}_{\tilde{E}_{a,t}}\mathbf{1}_{(E_{a,t}^{\theta})^c}\right] \\ &= \mathbb{E}[N_{a,\tau}] + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\mathbf{1}_{(a_t=a)}\mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}\left((E_{a,t}^{\theta})^c \mid \mathcal{F}_t\right)\right] \\ &= \mathbb{E}[N_{a,\tau}] + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\mathbf{1}_{(a_t=a)}\mathbf{1}_{\tilde{E}_{a,t}}\mathbb{P}\left(\mu(\theta_a(t) > \mu(\theta_a)+\Delta_a \mid Y_{a,t}\right)\right] \\ &\leq L_T + 1 + \max(N_\epsilon, N(\theta_a,F)) + \mathbb{E}\left[\sum_{t=\tau+1}^{T}\frac{1}{T}\right] \\ &\leq L_T + \max(N_\epsilon, N(\theta_a,F)) + 2.\end{aligned}$$

D Controling the Number of Optimal Plays: Outline Proof of Proposition 5

The proof of this proposition is quite detailed, and essentially the same as the proof given for Proposition 1 in [10], which we will sometimes refer to. However, in generalising to the case of exponential family bandits we show how to avoid the need to explicity calculate posterior probabilities that lead to Lemma 4 in [10]. While simplifying the proof we loose the ability to specify the constants explicitly, and so the analysis becomes asymptotic, but holds for every $b \in]0,1[$.

Sketch of the proof and key results Let τ_j be the occurrence of the j^{th} play of the optimal arm (with $\tau_0 := 0$). Let $\xi_j := (\tau_{j+1} - 1) - \tau_j$: this random variable measures the number of time steps between the j^{th} and the $(j+1)^{th}$ play of the optimal arm, and so $\sum_{a=2}^{K} N_{a,t} = \sum_{j=0}^{N_{1,t}} \xi_j$. We then upper bound $\mathbb{P}(N_{1,t} \leq t^b)$ as in [10]:

$$\mathbb{P}(N_{1,t} \leq t^b) \leq \mathbb{P}\left(\exists j \in \left\{0, .., \lfloor t^b \rfloor\right\} : \xi_j \geq t^{1-b} - 1\right) \leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}(\underbrace{\xi_j \geq t^{1-b} - 1}_{:=\mathcal{E}_j}) \tag{16}$$

We introduce the interval $\mathcal{I}_j = \{\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil\}$: on the event $\mathcal{E}_j$, $\mathcal{I}_j$ is included in $\{\tau_j, \tau_{j+1}\}$ and no draw of arm 1 occurs on $\mathcal{I}$. We also introduce for each arm $a \neq 1$ $d_a := \frac{\mu_1 - \mu_a}{2}$.

The idea of the rest of the analysis is based on the following remark. If on a subinterval $\mathcal{I} \subseteq [\tau_j, \tau_{j+1}[$ of size $f(t)$ arm 1 is not drawn and all the samples of the suboptimal arms fall below $\mu_2 + d_2 < \mu_1$, then for all $s \in \mathcal{I}$, $\mu(\theta_{1,s}) \leq \mu_2 + d_2$. On $\mathcal{I}$, the sequence $(\theta_{1,s})$ is i.i.d. with distribution π_{1,τ_j}, and hence,

$$\mathbb{P}(\forall s \in \mathcal{I},\ \mu(\theta_{1,s}) \leq \mu_2 + \delta) \leq \left(\mathbb{P}\left(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2\right)\right)^{f(t)}$$

At this point, an asymptotic result, telling that the posterior on θ_1 concentrates to a Dirac in θ_1 (the Bernstein-Von-Mises theorem, see [15]) , leads to

$$\mathbb{P}(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2) \underset{j \to \infty}{\to} 0.$$

Assuming that $\forall j$, $\mathbb{P}(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2) \neq 1$, we have shown the following Lemma, which plays the role of an asymptotic couterpart for Lemma 3 in [10].

Lemma 9. *There exists a constant $C = C(\pi_0) < 1$, such that for every (random) interval $\mathcal{I}$ included in $\mathcal{I}_j$ and for every positive function f, one has*

$$\mathbb{P}\left(\forall s \in \mathcal{I},\ \mu(\theta_{1,s}) \leq \mu_2 + \delta_2, \quad |\mathcal{I}| \geq f(t)\right) \leq C^{f(t)}.$$

Another key lemma is the following which generalizes Lemma 4 in [10]. The proof of this lemma is standard: it proceeds by conditioning on the event $\tilde{E}_{a,t}$[1] and applying Theorem 4, and Lemma 3.

Lemma 10. *For every $a \in A,\ \delta > 0$, there exist constants $C_a = C_a(\mu_a, \delta, F)$ and N such that for $t \geq N$,*

$$\mathbb{P}\left(\exists s \leq t, \exists a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a, N_{a,s} > C_a \ln(t)\right) \leq \frac{2(K-1)}{t^2}.$$

The rest of the proof proceeds by finding a subinterval of $\mathcal{I}_j$ on which all the samples of all the suboptimal arms indeed fall below the corresponding thresholds $\mu_a + d_a$. This is done exactly as in [10] and we recall the main steps of the proof below. Before that, we need to introduce the notion of *saturated*, suboptimal action.

Definition 11. *Let t be fixed. For any $a \neq 1$, an action a is said to be* saturated *at time s if it has been chosen at least $C_a \ln(t)$ times, i.e. $N_{a,t} \geq C_a \ln(t)$. We shall say that it is* unsaturated *otherwise. Furthermore at any time we call a choice of an unsaturated, suboptimal action an* interruption.

[1] Using $\tilde{E}_{a,t}$ in place of $E_{a,t}$ from [10] only changes slightly the constant C_a.

Step 1: Decomposition of $\mathcal{I}_j$ We want to study the process of saturation on the event $\mathcal{E}_j = \{\xi_j \geq t^{1-b} - 1\}$. We start by decomposing the interval $\mathcal{I}_j = \{\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil\}$ into K subintervals:

$$\mathcal{I}_{j,l} := \left\{ \tau_j + \left\lceil \frac{(l-1)(t^{1-b}-1)}{K} \right\rceil, \tau_j + \left\lceil \frac{l(t^{1-b}-1)}{K} \right\rceil \right\},\ l = 1, \dots, K.$$

Now for each interval $\mathcal{I}_{j,l}$, we introduce:

- $\mathcal{F}_{j,l}$: the event that by the end of the interval $\mathcal{I}_{j,l}$ at least l suboptimal actions are saturated;
- $n_{j,l}$: the number of interruptions during this interval.

We use the following decomposition to bound the probability of the event $\mathcal{E}_j$:

$$\mathbb{P}(\mathcal{E}_j) = \mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1}) + \mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1}) \tag{17}$$

Note that the quantities $\mathcal{E}_j,\ \mathcal{I}_{j,l},\ \mathcal{F}_{j,l}$ and $n_{j,l}$ all depend on t, however we suppress this dependency for notational convenience. However, we keep in mind that we bound the different probabilities for $t \geq N$, so that Lemma 10 applies.

Step 2: Bounding $\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1})$ On the event $\mathcal{E}_j \cap \mathcal{F}_{j,K-1}$, only saturated suboptimal arms are drawn on the interval $\mathcal{I}_{j,K}$. Using Lemma 10, we get

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \leq & \mathbb{P}(\{\exists s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) \leq \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
\leq & \mathbb{P}(\exists s \leq t, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a, N_{a,t} > C_a \ln(t)) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
\leq & \frac{2(K-1)}{t^2} + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K} : \mu(\theta_{1,s}) \leq \mu_2 + d_2\} \cap \mathcal{E}_j) \\
\leq & \frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{K}}.
\end{aligned}$$

for $0 < C < 1$ as in Lemma 9. The second last inequality comes from the fact that if arm 1 is not drawn, the sample $\theta_{1,s}$ must be smaller than some sample $\theta_{a,s}$ and therefore smaller than $\mu_2 + d_2$.

Step 3: Bounding $\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1})$ A similar argument to that employed in Step 2 can be used in an induction to show that for all $2 \leq l \leq K$, if t is larger than some deterministic constant $N_{\mu_1,\mu_2,b}$ specified in the base case,

$$\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,l-1}) \leq (l-2)\left(\frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{CK^2 \ln(t)}}\right)$$

We refer the reader to [10] for a precise description of the induction. For $l = K$ we then get

$$\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1}) \leq (K-2)\left(\frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{CK^2 \ln(t)}}\right). \tag{18}$$

Step 4: Conclusion Putting Steps 2 and 3 together we obtain that for $t \geq N_0 := \max(N, N_{\mu_1,\mu_2,b})$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j(t)) &\leq \frac{2(K-1)^2}{t^2} + C^{\frac{t^{1-b}-1}{K}} + (K-2)KC\ln(t)C^{\frac{t^{1-b}-1}{CK^2\ln(t)}}, \\
\mathbb{P}(N_{1,t} \leq t^b) &\leq \frac{2(K-1)^2}{t^{2-b}} + t^b C^{\frac{t^{1-b}-1}{K}} + (K-2)KCt^b\ln(t)C^{\frac{t^{1-b}-1}{CK^2\ln(t)}},
\end{aligned}$$

where we use 16. It then follows that

$$\sum_{t=1}^{\infty} \mathbb{P}(N_{1,t} \leq t^b) \leq N_0 + \sum_{t=N_0+1}^{\infty} \mathbb{P}(\mathcal{E}_j) = C_b = C_b(\pi_0, \mu_1, \mu_2, K) < \infty.$$